(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: **07056945 A**

(43) Date of publication of application: **03.03.95**

(51) Int. Cl **G06F 17/30**
G06F 12/00

(21) Application number: **05204351**

(22) Date of filing: **18.08.93**

(71) Applicant: **TOPPAN PRINTING CO LTD**

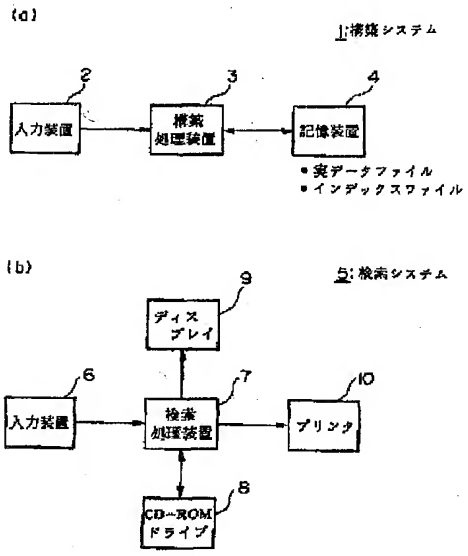
(72) Inventor: **UEKI HIROMI**
MAKINOUCHI KOUJI
HIRASAWA MICHIHIKO
NARA MASAHITO
HAMAYA GUNJI

(54) WHOLE SENSITIVE DATA BASE SYSTEM

(57) Abstract:

PURPOSE: To provide a whole sentence data base system where the size of an index file is set to be small and sufficient retrieval speed can be obtained.

CONSTITUTION: This system is provided with an input device 2 inputting document data, a construction processor 3 generating a Japanese sentence lowest layer table J2-1 and a Japanese sentence highest layer table J2-2, which have peculiar pseudo words whose character string length is more than '2', in terms of hierarchy based on document data supplied from the input device 2, a storage device 4 storing document data and the respective tables J2-1 and J2-2 as a real data file and an index file, an input device 6 inputting a retrieval character string and a retrieval processor 7 extracting the peculiar pseudo word which agrees with a retrieval pseudo word whose character string length constituting the retrieval character string becomes '2' from the index file and outputting document data corresponding to the retrieval character string to a display 9.



(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（A）**　(11)特許出願公開番号

特開平7－56945

(43)公開日　平成７年(1995)３月３日

(51)Int.Cl.[6]		識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ	17/30					
	12/00	５２０　Ａ	8944－5B			
			9194－5L	Ｇ０６Ｆ 15/ 40	３７０　Ａ	
			9194－5L	15/ 413	３１０　Ａ	

審査請求　未請求　請求項の数５　ＯＬ　（全 15 頁）

(21)出願番号　特願平5－204351

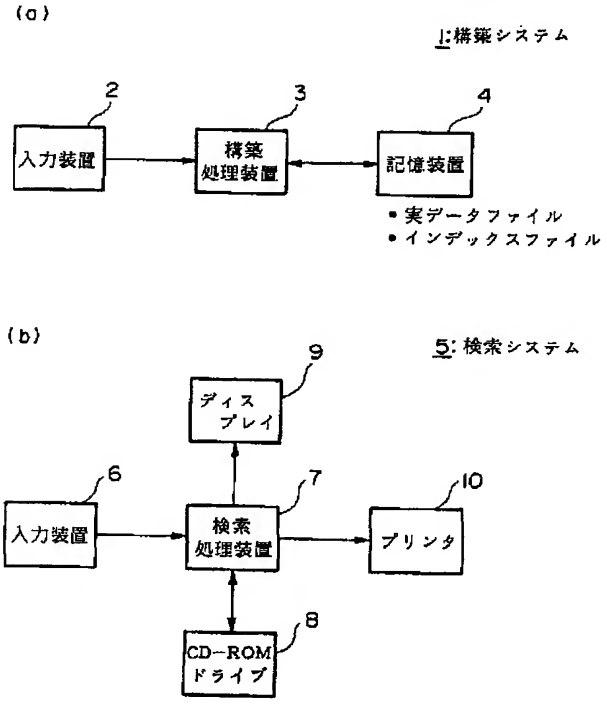
(22)出願日　平成５年(1993)８月18日

(71)出願人　000003193
凸版印刷株式会社
東京都台東区台東１丁目５番１号
(72)発明者　植木　広実
東京都台東区台東一丁目５番１号　凸版印刷株式会社内
(72)発明者　牧之内　浩二
東京都台東区台東一丁目５番１号　凸版印刷株式会社内
(72)発明者　平澤　道彦
東京都台東区台東一丁目５番１号　凸版印刷株式会社内
(74)代理人　弁理士　志賀　正武　（外２名）

最終頁に続く

(54)【発明の名称】　全文データベースシステム

(57)【要約】
【目的】　インデックスファイルのサイズを小とし、十分な検索速度を得ることができる全文データベースシステムを提供する。
【構成】　文書データを入力する入力装置２と、入力装置２から供給される文書データに基づいて、文字列長が２以上である固有疑似単語を有する和文最下層テーブルＪ２－１および和文最上層テーブルＪ２－２等を階層的に作成する構築処理装置３と、文書データおよび各テーブルＪ２－１，Ｊ２－２等を実データファイルおよびインデックスファイルとして記憶する記憶装置４と、検索文字列を入力する入力装置６と、検索文字列を構成する文字列長が２となる検索用疑似単語に一致する固有疑似単語を上記インデックスファイルから抽出し、検索文字列に対応する文書データをディスプレイ９へ出力する検索処理装置７とから構成される。



1

【特許請求の範囲】

【請求項１】　階層構造を有するインデックステーブル群を用いて、検索文字列に一致する文字列を文書データから抽出する全文データベースシステムであって、

前記文書データ中の各文字に連続するアドレスを付与するアドレス付与手段と、

前記文書データ中の各文字と後続する文字とで構成される合計ｋ文字（ｋは２以上）の疑似単語を作成し、各疑似単語の先頭文字列に付与される各アドレスを、対応する疑似単語の文字コード順にアドレステーブルへ記憶するアドレステーブル作成手段と、

固有の文字コードを有する疑似単語を固有疑似単語として前記インデックステーブル群中の最下層テーブルへ文字コード順に記憶するとともに、各固有疑似単語に前記アドレステーブル中の各アドレスを対応付ける最下層構築手段と、

前記インデックステーブル群中の最上層テーブルに記憶される固有疑似単語数が予め設定された数より大である場合、前記最上層テーブルを略均等に分割するように複数の固有疑似単語を抽出し、前記最上層テーブルの上層のテーブルへ前記複数の固有疑似単語を文字コード順に記憶する階層化手段とを具備することを特徴とする全文データベースシステム。

【請求項２】　前記インデックステーブル群は、合計ｋ文字未満の疑似単語に対応したテーブルをも有することを特徴とする請求項１に記載の全文データベースシステム。

【請求項３】　検索文字列に一致する文字列を文書データから抽出する全文データベースシステムであって、

連続するアドレスが付与された前記文書データ中の各文字、および当該各文字に後続する文字から構成される合計ｋ文字（ｋは２以上）の各疑似単語の先頭文字列に付与された各アドレスを、対応する疑似単語の文字コード順に記憶するアドレステーブルと、

階層構造を有するテーブル群であって、固有の文字コードを有する疑似単語を固有疑似単語として文字コード順に記憶するとともに、各固有疑似単語に前記アドレステーブル中の各アドレスが対応付けられる最下層テーブル、この最下層テーブルの上層に構築される複数の上層テーブルからなり、各上層テーブルは１段下層のテーブルを略均等に分割する位置に記憶された固有疑似単語を文字コード順に記憶する階層構造をなし、最上層テーブルに記憶される固有疑似単語の数が予め設定された数以下となるよう構成されるインデックステーブル群と、

検索文字列を入力する入力手段と、

該入力手段から供給される前記検索文字列をｋ文字単位に分割し、複数の検索用疑似単語を生成する分割手段と、

前記各検索用疑似単語と文字コードが同一である固有疑似単語を前記インデックステーブル群から抽出する抽出処理を行い、抽出された各固有疑似単語に対応する各アドレスの差から前記文書データ中で連続して存在する固有疑似単語の組を特定するとともに、前記組に対応するアドレス群に応じた文書データ中の文字列を出力する検索手段とを具備し、

前記検索手段は、前記検索用疑似単語と文字コードが一致する固有疑似単語が検索対象となるテーブル中に存在しない場合には、前記検索用疑似単語より小なる文字コードの固有疑似単語群から最も前記検索用疑似単語に近い文字コードの最小疑似単語と、この最小疑似単語の直後に記憶される最大疑似単語とを抽出するとともに、１段下層のテーブルを検索対象とし、当該テーブルにおいて、前記最小疑似単語に一致する文字コードの固有疑似単語と、前記最大疑似単語に一致する文字コードの固有疑似単語とに挟まれる範囲に対して前記抽出処理を施すことを特徴とする全文データベースシステム。

【請求項４】　階層構造を有するインデックステーブル群を用いて、検索文字列に一致する文字列を文書データから抽出する全文データベースシステムであって、

前記文書データ中の各文字に連続するアドレスを付与するアドレス付与手段と、

前記文書データ中の各文字と後続する文字とで構成される合計ｋ文字（ｋは２以上）の疑似単語を作成し、各疑似単語の先頭文字に付与される各アドレスを、対応する疑似単語の文字コード順にアドレステーブルへ記憶するアドレステーブル作成手段と、

固有の文字コードを有する疑似単語を固有疑似単語として前記インデックステーブル群中の最下層テーブルへ文字コード順に記憶するとともに、各固有疑似単語に前記アドレステーブル中の各アドレスを対応付ける最下層構築手段と、

前記インデックステーブル群中の最上層テーブルに記憶される固有疑似単語数が予め設定された数より大である場合、前記最上層テーブルを略均等に分割するように複数の固有疑似単語を抽出し、前記最上層テーブルの上層のテーブルへ前記複数の固有疑似単語を文字コード順に記憶する階層化手段と、

前記検索文字列を入力する入力手段と、

該入力手段から供給される前記検索文字列をｋ文字単位に分割し、複数の検索用疑似単語を生成する分割手段と、

前記インデックステーブル群から前記検索用疑似単語と文字コードが同一である固有疑似単語を抽出する抽出処理を行い、抽出された各固有疑似単語に対応する各アドレスの差から前記文書データ中で連続して存在する固有疑似単語の組を特定するとともに、前記組に対応するアドレス群に応じた文書データ中の文字列を出力する検索手段とを具備し、

前記検索手段は、前記検索用疑似単語と文字コードが一致する固有疑似単語が検索対象となるテーブル中に存在

しない場合には、前記検索用疑似単語より小なる文字コードの固有疑似単語群から最も前記検索用疑似単語に近い文字コードの最小疑似単語と、この最小疑似単語の直後に記憶される最大疑似単語とを抽出するとともに、１段下層のテーブルを検索対象とし、当該テーブルにおいて、前記最小疑似単語に一致する文字コードの固有疑似単語と、前記最大疑似単語に一致する文字コードの固有疑似単語とに挟まれる範囲に対して前記抽出処理を施すことを特徴とする全文データベースシステム。

【請求項５】　前記インデックステーブル群は、合計ｋ文字未満の疑似単語に対応したテーブルをも有し、前記検索手段は前記検索文字列の長さに応じて前記検索対象とするテーブルを変更することを特徴とする請求項３または４に記載の全文データベースシステム。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は電子出版等に用いて好適な全文データベースシステムに関する。

【０００２】

【従来の技術】文書データから所定の文字列を検索する手法には、大別して、キーワード検索と全文検索とがある。キーワード検索では、キーワード登録者が、文書データを例えば、文章毎に細分化し、各文書データに適合するキーワードを対応させる。この対応関係はインデックスファイルに記憶され、当該ファイルにおいて、各キーワードは例えば、文字コード順に木構造をなすように記憶される。

【０００３】このような構成において、検索操作者が抽出したい文書データに対応するキーワードを入力することにより検索処理が為される。具体的には、入力キーワードに一致するキーワードがインデックスファイルから抽出され、このキーワードに対応する文書データが出力される。

【０００４】一方、全文検索では、文書データはシーケンシャルファイル（全文データベース）として記憶される。この文書データに対する検索処理は、検索操作者が、抽出したい文書データに含まれる文字列を入力することにより為される。具体的には、入力された検索文字列と文書データに含まれる全ての文字列とが比較され、一致する文字列が含まれる文書データが出力される。なお、検索文字列の文字数や適用するパターンマッチングの手法によっては、文書データに含まれる全ての文字列を検索文字列と比較する必要はない。各種パターンマッチングの手法については、公知であるので、その説明を省略する。

【０００５】

【発明が解決しようとする課題】一般に、データベース検索を必要とする電子出版の分野において、取り扱われる文書データの容量は、極めて大きなものとなる。例えば、ＣＤ－ＲＯＭは数百ＭＢ（メガバイト）の記憶容量を有しており、通常、数十～数百ＭＢの文書データが記憶される。このように、極めて大容量の文書データに対して、前述した全文検索を適用すると、公知のいかなるパターンマッチングの手法を用いても、パーソナルコンピュータ程度の処理能力では、十分な応答時間を達成することができないという欠点がある。

【０００６】また、前述したキーワード検索を適用しようとしても、ＣＤ－ＲＯＭに記憶される文書データは、データベース化を前提として作成されていない為に、文章毎に細分化する作業や、適合するキーワードを作成する作業が極めて困難になるという問題がある。さらに、キーワード登録者と検索操作者は別人であるために、検索操作者が、作成されたキーワードを有効に活用することが困難であるという欠点もある。

【０００７】また、使用者が必要とする情報は、「経済」、「製法」、「演算」等の数文字の文字列により特定できる場合が多く、これらの文字列の中には、キーワードとして登録しにくいものが含まれることがある。すなわち、登録されたキーワード以外の文字列による検索が必要となる場合がある。こうした場合に対応できるように、例えば、２文字からなる文字列を全てキーワードとすることも考えられる。

【０００８】しかしながら、日本語の場合、約７千種の文字が存在するために、２文字の順列は約４千９百万という膨大な数となる。この膨大な数のキーワードから特定のキーワードを抽出する処理は極めて高負荷となり、木構造を用いても、パーソナルコンピュータ程度の処理能力では実用にならない。また、検索の為のインデックスファイルも巨大（例えば、数百ＭＢ）なものとなり、文書データの為の記憶領域が狭くなってしまう。

【０００９】本発明は、このような背景に鑑みて為されたもので、インデックステーブル群のサイズを小とし、十分な検索速度を得ることができる全文データベースシステムを提供することを目的とする。

【００１０】

【課題を解決するための手段】本発明による全文データベースシステムは、階層構造を有するインデックステーブル群を用いて、検索文字列に一致する文字列を文書データから抽出する全文データベースシステムであって、前記文書データ中の各文字に連続するアドレスを付与するアドレス付与手段と、前記文書データ中の各文字と後続する文字とで構成される合計ｋ文字（ｋは２以上）の疑似単語を作成し、各疑似単語の先頭文字列に付与される各アドレスを、対応する疑似単語の文字コード順にアドレステーブルへ記憶するアドレステーブル作成手段と、固有の文字コードを有する疑似単語を固有疑似単語として前記インデックステーブル群中の最下層テーブルへ文字コード順に記憶するとともに、各固有疑似単語に前記アドレステーブル中の各アドレスを対応付ける最下層構築手段と、前記インデックステーブル群中の最上層

テーブルに記憶される固有疑似単語数が予め設定された数より大である場合、前記最上層テーブルを略均等に分割するように複数の固有疑似単語を抽出し、前記最上層テーブルの上層のテーブルへ前記複数の固有疑似単語を文字コード順に記憶する階層化手段とを具備することを特徴としている。

【００１１】

【作用】上記構成によれば、アドレス付与手段が、文書データ中の各文字に連続するアドレスを付与し、アドレステーブル作成手段が、前記文書データ中の各文字と後続する文字とで構成される合計ｋ文字（ｋは２以上）の疑似単語を作成し、各疑似単語の先頭文字列に付与される各アドレスを、対応する疑似単語の文字コード順にアドレステーブルへ記憶する。そして、最下層構築手段が、固有の文字コードを有する疑似単語を固有疑似単語として前記インデックステーブル群中の最下層テーブルへ文字コード順に記憶するとともに、各固有疑似単語に前記アドレステーブル中の各アドレスを対応付ける。さらに、階層化手段が、前記インデックステーブル群中の最上層テーブルに記憶される固有疑似単語数が予め設定された数より大である場合、前記最上層テーブルを略均等に分割するように複数の固有疑似単語を抽出し、前記最上層テーブルの上層のテーブルへ前記複数の固有疑似単語を文字コード順に記憶する。インデックステーブル群は、このような階層構造を有する為に、検索時において、十分な検索速度が得られる。また、インデックステーブル群中の各テーブルに記憶される疑似単語は、固有の文字コードを有する固有疑似単語であるために、インデックステーブル群のサイズが小となる。

【００１２】

【実施例】以下、図面を参照して、本発明の一実施例について説明する。

（１）構築システム１の構成

図１は本発明の一実施例による全文データベースシステムの概略構成を示す図であり、図１（ａ）は全文データベースを構築する構築システム１の概略構成を示すブロック図である。この構築システム１は、データベースの提供者（もしくは編集者）に使用されることが想定される。電子出版においては、その提供者が当該システム１＊により構築された全文データベースを、ＣＤ－ＲＯＭ等に記憶させる。

【００１３】図１（ａ）において、２はキーボード等の入力装置であり、オペレータにより入力される文書データを構築処理装置３（後述する）へ供給する。構築処理装置３は、入力装置２から供給される文書データに所定の処理を施して、記憶装置４（後述する）へ供給するとともに、当該文書データに対応するインデックスファイルを作成し、全文データベースを構築する。この全文データベース構築処理の内容については、後に詳述する。

【００１４】記憶装置４は、例えば、数百ＭＢの容量を有するハードディスクからなり、構築処理装置３から供給される文書データを実データファイルとして記憶するとともに、構築処理装置３にて作成されるインデックスファイルを記憶する。ここで、実データファイル中の文書データの一例を図４に示す。この図に示す文書データには、キーワード”東京都”および”京都”を識別する為の位置マーク’＠’が付加されている。

【００１５】また、インデックスファイルに含まれる各種テーブルの一例を図１２に示す。図１２は図４に示す文書データの和文範囲ＪＡに対するインデックステーブル（以後、和文テーブルと称す）の構成例を示す図であり、この図において、Ｊ２－１は和文最下層テーブル、Ｊ２－２は和文最下層テーブルＪ２－１の上層テーブルとなる和文最上層テーブル、Ｊ１は１文字テーブルであり、和文最下層テーブルＪ２－１の上層テーブルとなる。また、ＣＡＴはレコードＲ１を複数有する文字アドレステーブルである。

【００１６】和文最下層テーブルＪ２－１はレコードＲ２を複数有し、各レコードＲ２には、文書データの和文範囲ＪＡ（図４参照）中に存在する２文字の文字列（以後、疑似単語と称す）が格納されている。これらの疑似単語は、ユニーク（同一綴りのものが無い）であり、文字コードＣＣ順にソートされている（以後、ユニークな疑似単語を固有疑似単語と称す）。ここで、疑似単語の文字コードＣＣは、当該疑似単語の先頭文字の文字コードをＣ１、末尾文字の文字コードをＣ２とすると、例えば、以下に示す計算式（１）により算出される。

$$CC = C1 \times m + C2 \quad \cdots (1)$$

文書データ中に現れる文字種は約７千種であるので、この文字種以上の数「ｍ（例えば、８千）」をＣ１に乗ずることにより、ＣＣは疑似単語固有の文字コードとなる。

【００１７】また、各レコードＲ２は、当該レコードＲ２に格納される固有疑似単語の文書データ中での存在数を表す「サイズ」を有し、文字アドレステーブルＣＡＴ中の所定のレコードＲ１に対応付けられている。ここで、レコードＲ２が対応付けられるレコードＲ１は、当該レコードＲ２が有する固有疑似単語の先頭文字の位置を表す「文字アドレス」を有する。

【００１８】なお、和文最下層テーブルＪ２－１において、サイズが複数であるレコードＲ２には、複数のレコードＲ１が対応付けられる。具体的には、サイズが複数であるレコードＲ２に、当該レコードＲ２が有する固有疑似単語に対応する複数の文字アドレスのうち、最小の文字アドレスを有するレコードＲ１が対応付けられ、このレコードＲ１に続いて、最小でない文字アドレスを有するレコードＲ１が昇順に整列される。

【００１９】また、和文最上層テーブルＪ２－２はレコ

ードＲ３を複数有し、各レコードＲ３には、和文最下層テーブルＪ２－１中の固有疑似単語が格納される。なお、このレコードＲ３の数は、レコードＲ２に較べて極めて少ない。また、各レコードＲ３は、固有疑似単語の文字コード順にソートされており、レコードＲ３が有する固有疑似単語と同一の固有疑似単語を有するレコードＲ２に対応付けられる。この際、和文最下層テーブルＪ２－１において、各レコードＲ３に対応付けられる各レコードＲ２間の距離、すなわち、当該レコードＲ２間に存在するレコードＲ２の数が、予め設定された数（例えば、９９）となるように、各レコードＲ３に格納される固有疑似単語が設定される。

【００２０】１文字テーブルＪ１はレコードＲ４を複数有し、各レコードＲ４は、文書データの和文範囲ＪＡに出現する全種類の文字を格納する。これらの文字はユニークであり、各レコードＲ４は文字コード順にソートされている。また、各レコードＲ４は、和文最下層テーブルＪ２－１中の所定のレコードＲ２に対応付けられており、対応付けられるレコードＲ４およびレコードＲ２が有する文字および固有疑似単語の先頭文字は一致する。ここで、１つのレコードＲ４が対応付けられるべきレコードＲ２が複数である場合には、当該レコードＲ４は最も先頭にあるレコードＲ２に対応付けられる。この際、レコードＲ４が有するサイズは、対応付けるべき固有疑似単語の数となる。

【００２１】次に、文書データ（図４参照）の欧文範囲ＥＡに対応するインデックステーブル（以後、欧文テーブルと称す）の一例を図７および図９に示す。欧文テーブルは、図７に示すようなワードアドレステーブルＷＡＴ，ワードリストＷＬと、図９に示すような仮想アドレステーブルＶＡＴ，欧文基本テーブルＥＢＴとからなる。図７に示すように、ワードアドレステーブルＷＡＴはレコードＲ５を複数有し、各レコードＲ５には、欧文範囲ＥＡに存在する全ての「ワード（単語）」に対応するワード単位のアドレス（以後、ワードアドレスと称す）が格納される。

【００２２】また、ワードリストＷＬは、レコードＲ６を複数有し、各レコードＲ６は、欧文範囲ＥＡに存在するユニークなワード（以後、固有ワードと称す）を１つずつ有し、各固有ワードの文字コード順にソートされている。また、各レコードＲ６には、固有の「ユニーク符号」が割り当てられ、各固有ワードを構成する各文字には、固有ワード内の先頭からの位置を示す「ワード内アドレス」が付与される。

【００２３】また、各レコードＲ６は、自身が有する固有ワードが文書データ中で出現する頻度を表す「サイズ」を有し、ワードアドレステーブルＷＡＴ中の所定のレコードＲ５に対応付けられる。なお、サイズが複数であるレコードＲ６には、複数のレコードＲ５が対応付けられる。具体的には、サイズが複数であるレコードＲ６に、当該レコードＲ６が有する固有ワードに対応する複数のワードアドレスのうち、最小のワードアドレスを有するレコードＲ５が対応付けられ、このレコードＲ５に続いて、最小でない文字アドレスを有するレコードＲ５が昇順に整列される。

【００２４】また、図９に示す欧文基本テーブルＥＢＴは、レコードＲ７を複数有し、各レコードＲ７には、ワードリストＷＬ（図７参照）中の各固有ワードに基づいて生成される２文字単位の固有疑似単語が１つずつ格納される。各レコードＲ７は、固有疑似単語の文字コード順にソートされている。各レコードＲ７は、格納された固有疑似単語のワードリストＷＬ中での出現頻度を表す「サイズ」を有し、仮想アドレステーブルＶＡＴ中の所定のレコードＲ８に対応付けられる。

【００２５】仮想アドレステーブルＶＡＴはレコードＲ８を複数有し、各レコードＲ８には、ユニーク符号とワード内アドレスとの組である「仮想アドレス」が格納される。この仮想アドレスは、欧文基本テーブルＥＢＴ中の固有疑似単語に対応するものである。ここで、レコードＲ７のサイズが複数である場合には、当該レコードＲ７に複数のレコードＲ８が対応付けられる。この対応付けの具体的内容は、図１２に示す和文最下層テーブルＪ２－１のレコードＲ２と、文字アドレステーブルＣＡＴのレコードＲ１との対応付けと同様であるので、その説明を省略する。

【００２６】（２）検索システム５の構成

一方、図１（ｂ）は構築システム１により構築された全文データベースに対して、各種の検索処理を行う検索システム５の構成を示す図である。この検索システム５は、データベースのユーザーによって用いられることが想定される。電子出版においては、そのユーザーが当該システム５を用いて、ＣＤ－ＲＯＭ等に記憶された全文データベースに対して各種の検索処理を行う。

【００２７】図１（ｂ）において、６はキーボード等の入力装置であり、ユーザー（使用者）により入力される指示に応じた指示データを検索処理装置７（後述する）へ供給する。検索処理装置７は、一般的なパーソナルコンピュータであり、図示せぬＣＰＵ、ＲＯＭ、ＲＡＭおよび各種Ｉ／Ｏインタフェースを有する。この検索処理装置７はＲＡＭに記憶される検索プログラムを実行し、入力装置６から供給される指示データに応じた検索処理を行う。この検索処理の内容は後に詳述する。

【００２８】８はＣＤ－ＲＯＭドライブであり、検索処理装置７に制御され、挿入されるＣＤ－ＲＯＭに記憶された情報を読み取る。９は検索処理装置７から供給される表示データに応じて、検索メニューや検索結果等を表示するディスプレイ、１０はプリンタであり、検索処理装置７から供給される出力データに応じて、検索結果を出力する。

【００２９】（３）全文データベース構築処理

次に、構築処理装置３（図１（ａ）参照）がＲＡＭに記憶されたプログラムを実行して行う全文データベース構築処理について、以下に説明する。ここでは、和文と欧文が混在した文書データ（図３参照）をデータベース化する場合について説明する。

【００３０】まず、全文データベース構築に先だって、図１（ａ）に示す構築システム１において、入力装置２から文書データが入力される。この文書データは、構築処理装置３を介して記憶装置４に供給され、ここで記憶される。以下に説明する各処理は、記憶装置４に記憶された文書データに対して為される。文書データの入力処理が終了し、入力装置２から所定の指示データが入力されると、構築処理装置３は、図２のフローチャートに表されるプログラムを実行する。

【００３１】まず、ステップＳＡ１では、文書データ（図３参照）に位置マークを付加する。この位置マークとは、任意のキーワードの前後に挿入される特定の文字であり、例えば、’＠’のように、文書データ中に存在しない記号を用いる。図４は、上述したマーク付加処理が行われた後の文書データを示す図であり、この図において、”東京都”および”京都”という文字列（キーワード）の前後には位置マーク’＠’が挿入されている。

【００３２】上述したマーク付加処理を行わない場合、”京都”という地名を全文検索すると、”京都”はもちろん、”東京都”まで抽出してしまう。マーク付加処理は、このような無意味な抽出を避ける為に行われる処理であり、全文検索において、検索文字列として”＠京都＠”という文字列を入力すると、”＠京都＠”のみが抽出され、”＠東京都＠”は抽出されないという結果を得ることができる。

【００３３】次に、ステップＳＡ２（図２参照）では、文書データにアドレスが付与される。図５に示すように、アドレスには文字アドレスとワードアドレスがあり、文字アドレスは文書データ全体に付与され、ワードアドレスはアルファベットや数字等が連続する欧文範囲ＥＡに付与される。例えば、図５の和文範囲ＪＡにおいて、先頭文字’多’の文字アドレスは「１」、それに続く文字’角’の文字アドレスは「２」となり、欧文範囲ＥＡにおいて、先頭文字’ｗ’の文字アドレスは「３１７」となる。また、欧文範囲ＥＡにおいて、先頭のワード”ｗｏｒｌｄ”のワードアドレスは「１」、それに続くワード”ｗｉｄｅ”のワードアドレスは「２」となる。

【００３４】後述する各処理において、欧文範囲ＥＡにはワード単位の処理が行われるため、欧文範囲ＥＡにおいて、文字アドレスを記憶する必要はない。しかしながら、和文範囲ＪＡとの位置関係を把握するために、欧文範囲ＥＡの最初および最後の文字アドレスを記憶装置４（図１（ａ）参照）に記憶する。これらの文字アドレス間の文書データは、後述する各処理において、欧文範囲ＥＡとみなされ、ワード単位の処理を施される。

【００３５】次に、ステップＳＡ３（図２参照）では、欧文範囲ＥＡのワードリストＷＬを作成する。まず、図５の文書データの欧文範囲ＥＡに出現する全てのワードを抽出し、図６に示すように、欧文対照テーブルＣＴを作成する。次に、欧文対照テーブルＣＴ中の各レコードＲ１１を、各ワードの文字コード順およびワードアドレス順にソートする。すると、同一のワード（例えば、図６中のワード”ｗｏｒｌｄ”参照）を有する複数のレコードＲ１１が隣接する。

【００３６】次に、ソートされた各レコードＲ１１からユニークな固有ワードを抽出し、ワードリストＷＬ（図７参照）の各レコードＲ６に格納するとともに、ワードアドレスを有するレコードＲ５を複数作成し、ワードアドレステーブルＷＡＴを作成する。ワードリストＷＬの各レコードＲ６に設けられるポインタは、ワードアドレステーブルＷＡＴ中のレコードＲ５を指し示す。

【００３７】この際、ポインタにより対応付けられるレコードＲ６の固有ワードとレコードＲ５のワードアドレスとは、欧文対照テーブルＣＴにおいて同一レコードＲ１１内に格納されていたもの同士となる。また、欧文対照テーブルＣＴにおいて、同一のワードを有するレコードＲ１１が複数存在していた場合、そのワードに一致する固有ワードを有するワードリストＷＬ中のレコードＲ６には、当該ワードの欧文対照テーブルＣＴ内での出現数がサイズとして格納される。例えば、ワード”ｈｅｌｐ”は、欧文対照テーブルＣＴ（図６参照）中に２つ出現するので、ワードリストＷＬの固有ワード”ｈｅｌｐ”を有するレコードＲ６のサイズは「２」となる。

【００３８】さらに、ワードリストＷＬの各レコードＲ６には、固有のユニーク符号「Ａ」，「Ｂ」，「Ｃ」，・・・が付与され、各レコードＲ６の固有ワードを構成する文字には、ワード内アドレスが付与される。例えば、固有ワード”ｃａｎ”に付与されるユニーク符号は「Ａ」であり、固有ワードを構成する文字’ｃ’に対するワード内アドレスは「１」である。こうして作成されたワードリストＷＬは、記憶装置４（図１（ａ）参照）に記憶される。

【００３９】次に、ステップＳＡ４（図２参照）では、文書データまたはワードリストＷＬから疑似単語を抽出する。これに続くステップＳＡ５では、抽出された疑似単語を用いて、和文最下層テーブルＪ２－１および欧文基本テーブルＥＢＴを作成する。これらの抽出処理および作成処理は、文書データの形式により異なる為、以下、欧文範囲ＥＡと和文範囲ＪＡとに分けて説明する。

【００４０】Ａ：欧文範囲ＥＡに対する処理

欧文範囲ＥＡにおいて、まず、ワードリストＷＬ（図７参照）から、文字列長が「２」である疑似単語を抽出する。この抽出処理は、各固有ワードの先頭から末尾にかけて行われ、例えば、固有ワード”ｃａｎ”からは、”

ｃａ”，”ａｎ”という疑似単語が抽出される。こうして抽出された複数の疑似単語は、図８に示すような構成の疑似単語テーブルＰＷＴの各レコードＲ９に格納される。

【００４１】疑似単語テーブルＰＷＴにおいて、疑似単語が格納されたレコードＲ９は、当該レコードＲ９が有する疑似単語の抽出元の固有ワードに付与されたユニーク符号と、その疑似単語の先頭文字のワード内アドレスとから構成される「仮想アドレス」を有する。例えば、疑似単語”ａｎ”を有するレコードＲ９は、抽出元の固有ワード”ｃａｎ”に付与されたユニーク符号「Ａ」と、疑似単語”ａｎ”の先頭文字’ａ’のワード内アドレス「２」とから構成される仮想アドレス「Ａ－２」を有する。

【００４２】そして、ワードリストＷＬ作成時と同様に、疑似単語テーブルＰＷＴ内の各レコードＲ９を各疑似単語の文字コード順にソートし、ユニークな固有疑似単語を抽出する。ここで、抽出された固有疑似単語は、図９に示す欧文基本テーブルＥＢＴに格納される。また、疑似単語テーブルＰＷＴ内の仮想アドレスのみで構成される仮想アドレステーブルＶＡＴを作成する。図９に示すように、欧文基本テーブルＥＢＴの各レコードＲ７に設けられるポインタは、仮想アドレステーブルＶＡＴ中の対応するレコードＲ８を指し示す。

【００４３】また、ワードリストＷＬ作成時と同様に、疑似単語テーブルＰＷＴにおいて、同一の疑似単語が複数存在していた場合、その疑似単語と同一の固有疑似単語を有するレコードＲ７には、その疑似単語の出現数が「サイズ」として格納される。こうして、図９に示す欧文基本テーブルＥＢＴが作成される。

【００４４】Ｂ：和文範囲ＪＡに対する処理

和文範囲ＪＡにおいては、まず、文書データ（図５参照）から、文字列長が「２」である疑似単語を抽出する。この抽出処理は、和文範囲ＪＡの先頭から末尾にかけて行われ、例えば、図５の文書データからは、順に”多角”，”角経”という疑似単語が抽出される。抽出された疑似単語は、図１０に示すような和文疑似単語テーブルＪＰＴの各レコードＲ１０に格納される。

【００４５】和文疑似単語テーブルＪＰＴにおいて、疑似単語が格納されたレコードＲ１０は、当該疑似単語の先頭文字の「文字アドレス」を有する。例えば、疑似単語”多角”を有するレコードＲ１０は、疑似単語の先頭文字’多’の文字アドレス「１」を有する。そして、ワードリストＷＬや疑似単語テーブルＰＷＴ作成時と同様に、各レコードＲ１０を疑似単語の文字コード順にソートし、ユニークな固有疑似単語を抽出する。

【００４６】ここで、抽出された固有疑似単語は、図１１に示す和文最下層テーブルＪ２－１に格納される。また、和文疑似単語テーブルＪＰＴ内の文字アドレスのみで構成される文字アドレステーブルＣＡＴを作成する。図１１に示すように、和文最下層テーブルＪ２－１の各レコードＲ２に設けられるポインタは、文字アドレステーブルＣＡＴ中の対応するレコードＲ１を指し示す。

【００４７】また、和文疑似単語テーブルＪＰＴ（図１０参照）において、同一の疑似単語が複数存在していた場合、その疑似単語に一致する固有疑似単語を有するレコードＲ２には、当該疑似単語の文書データ中での出現数が「サイズ」として格納される。こうして、図１１に示す和文最下層テーブルＪ２－１が作成される。上述したように、欧文基本テーブルＥＢＴおよび和文最下層テーブルＪ２－１が作成されると、処理はステップＳＡ６へ進む。

【００４８】ステップＳＡ６において、最上層の和文テーブル（最初は和文最下層テーブルＪ２－１）中のレコード数が所定数ｎ（例えば、ｎ＝５００）以上であるか否かが判断される。この判断が「Ｙｅｓ」であれば、処理はステップＳＡ７へ進み、「Ｎｏ」であれば、処理はステップＳＡ８へ進む。

【００４９】上記ステップＳＡ６での判断により、必要に応じて、和文テーブルが階層化されるのだが、ここで、当該階層化を行う理由を説明する。前述したように、和文に用いられる文字種は約７千種と多く、文書データから抽出される固有疑似単語の種類、すなわち、和文最下層テーブルＪ２－１のレコードＲ２の数は極めて大となる。後述する検索処理は、検索文字列に対応する固有疑似単語を抽出する処理を繰り返して行われるため、検索対象となるテーブル（例えば、和文最下層テーブルＪ２－１）のレコードＲ２の数が多いと、所定の時間内に検索処理を終了することができない。

【００５０】ここで、和文最下層テーブルＪ２－１のレコードＲ２の数が多い場合には、図１２に示すように、上層のテーブル（和文最上層テーブルＪ２－２）を作成し、上層のテーブルで検索文字列に対応する固有疑似単語を抽出できなかった場合には、下層テーブル（和文最下層テーブルＪ２－１）の所定の範囲で、固有疑似単語を抽出するようにする。すると、比較すべき固有疑似単語の数が減少し、所定の応答時間で検索処理を行うことができる。

【００５１】なお、欧文基本テーブルＥＢＴに関して上記階層化を行わないのは、当該テーブルＥＢＴ中の各固有疑似単語は、文字種の少ないアルファベットや数字等の組み合わせであるために、そのレコード数は、和文最下層テーブルＪ２－１のレコード数に較べて極めて少なく（同一綴りの疑似単語が多い）、欧文基本テーブルＥＢＴのみでも十分な応答時間を得ることができるからである。

【００５２】ステップＳＡ７は、ステップＳＡ６での判断が「Ｙｅｓ」となった場合の処理であり、ここでは、既に作成された和文テーブルに対して上層の和文テーブルを作成する。例えば、図１０の和文最下層テーブルＪ

２－１に対して、図１１に示すように、和文最上層テーブルＪ２－２を作成する。この和文最上層テーブルＪ２－２の各レコードＲ３には、”＠東”や”経常”等の和文最下層テーブルＪ２－１から抽出された固有疑似単語が格納される。また、各レコードＲ３は「ポインタ」を有し、和文最下層テーブルＪ２－１中の同一固有疑似単語を有するレコードＲ２に対応付けられる。

【００５３】なお、各レコードＲ３が有する固有疑似単語は、隣接するレコードＲ３に対応する各レコードＲ２間の距離（２つのレコードＲ２に含まれるレコードＲ２の数）が、例えば、９９となるように抽出される。この距離は和文最下層テーブルＪ２－１のレコード数に応じて設定される。そして、処理はステップＳＡ６に戻る。こうして、最上層の和文テーブルのレコード数が所定数ｎ未満となるまで、上述した階層化処理が行われる。図１２に示す例では、和文最上層テーブルＪ２－２のレコード数は所定数ｎ（例えば、ｎ＝５００）未満となるので、和文テーブルの階層は２段となる。

【００５４】ステップＳＡ８は、最上層の和文テーブルのデータ数が所定数ｎ未満となり、ステップＳＡ６での判断が「Ｎｏ」となる場合の処理であり、和文最下層テーブルＪ２－１に対する１文字テーブルＪ１が作成される。和文には漢字が用いられるため、「山」や「川」等の１文字の検索文字列による検索が行われる場合がある。こうした１文字検索をも所定の応答速度で実現する為に、１文字テーブルＪ１が作成される。

【００５５】１文字テーブルＪ１の作成過程を以下に説明する。まず、図１２に示す和文最下層テーブルＪ２－１から各固有疑似単語の先頭文字をレコード順に抽出する。各レコードＲ２は、既に文字コード順にソートされている為、同一の文字が連続して抽出される。次に、抽出された文字群からユニークな文字を抽出し、抽出元の文字群に含まれる同一文字の数（サイズ）とともに、１文字テーブルＪ１の各レコードＲ４に格納する。

【００５６】また、各レコードＲ４はポインタを有し、和文最下層テーブルＪ２－１内の抽出元レコードＲ２に対応付けられる。こうして、１文字テーブルＪ１が作成される。そして、例えば、図７のワードアドレステーブルＷＡＴ，ワードリストＷＬと、図９の仮想アドレステーブルＶＡＴ，欧文基本テーブルＥＢＴと、図１２の文字アドレステーブルＣＡＴ，和文最下層テーブルＪ２－１，和文最上層テーブルＪ２－２，１文字テーブルＪ１と、欧文範囲ＥＡの最初および最後の文字アドレスとが、インデックスファイルとして、記憶装置４に記憶される。また、図５に示すような文書データが実データファイルとして記憶装置４に記憶され、全文データベースが構築される。

【００５７】（４）全文検索処理

次に、上述した過程を経て構築された全文データベースに対して、検索処理装置７（図１（ｂ）参照）が行う全文検索処理について、図面を参照して説明する。図１３，図１４は検索処理装置７のＲＡＭに予め記憶される全文検索プログラムのフローチャートである。まず、検索システム５において、ＣＤ－ＲＯＭドライブ８に、全文データベースが記憶されたＣＤ－ＲＯＭが挿入され、入力装置６から所定の指示データが供給されと、検索処理装置７はステップＳＢ１を実行する。

【００５８】ステップＳＢ１では、所定の表示データをディスプレイ９へ供給し、例えば、図１５に示す検索メニューを表示させる。ユーザーは、表示された検索メニューに応じて、入力装置６を操作し、後述する検索モード、順位モードおよび指定距離等を設定する。検索モードには１つの検索文字列を検索する通常検索モードの他に、複数の検索文字列を文脈上の関係を意識して検索する文脈意識モードがあり、ユーザーは入力装置６を操作して入力フィールド１１へ所定の文字を入力し、どちらかのモードを選択する。

【００５９】分脈意識モードを選択した場合、ユーザーは複数の検索文字列間の前後関係を意識するか否か（順位モード）を指定する必要がある。また、分脈意識モードでは、複数の検索文字列間の距離（先頭文字アドレスの差）の上限を指定する必要がある。したがって、ユーザーは入力フィールド１２に順位モードを指定する文字を入力し、入力フィールド１３に距離の上限を表す数値（指定距離）を入力する。

【００６０】次に、ステップＳＢ２では、ユーザーが入力装置６を操作し、図１４の文字列入力フィールド１４、あるいは文字列入力フィールド１５へ検索文字列を入力する。そして、入力装置６から所定の指示データが供給されると、検索処理装置７は、検索メニュー上の各入力フィールド１１～１５に入力された各種のデータを読み取り、これらのデータをＲＡＭに記憶する。そして、処理はステップＳＢ３へ進む。

【００６１】ステップＳＢ３では、検索文字列の各文字に文字アドレスを付与し、検索文字列を２文字単位に分割して、複数の検索用疑似単語を抽出する。例えば、図１７に示す”経営危機”という検索文字列からは”経営”と”危機”という検索用疑似単語が抽出される。検索文字列が２文字以下の長さであれば、上記抽出処理は行われない。

【００６２】次に、ステップＳＢ４では、各検索用疑似単語に一致する固有疑似単語を有するレコードを、記憶装置４に記憶されたインデックスファイルから検索する。この検索処理は各検索用疑似単語の文字コードＳＣと、インデックスファイル中の各固有疑似単語の文字コードＶＣとを比較することにより行われる。ここで、検索処理に使用されるテーブルは、欧文範囲ＥＡでの検索では欧文基本テーブルＥＢＴ、和文範囲ＪＡでの検索では最上層の和文テーブル（例えば、和文最上層テーブルＪ２－２）あるいは１文字テーブルＪ１となる。そし

て、ステップＳＢ５では、上記検索処理が全ての検索用疑似単語に対して完了したか否かを判断する。この判断が「Ｎｏ」の場合はステップＳＢ６へ、逆に「Ｙｅｓ」の場合はステップＳＢ９へ処理が進む。

【００６３】ステップＳＢ６は、検索用疑似単語に対する検索処理が完了しなかった場合の処理であり、検索対象となっている和文テーブルが最下層のテーブル（例えば、和文最下層テーブルＪ２－１）であるか否かを判断する。この判断が「Ｎｏ」の場合はステップＳＢ７へ、逆に「Ｙｅｓ」の場合はステップＳＢ１８（図１４参照）へ処理が進む。

【００６４】ステップＳＢ７は、検索対象となっている和文テーブルが、さらに下層のテーブルを有する場合の処理である。ここでは、検索対象となっている和文テーブル（例えば、和文最上層テーブルＪ２－２）において、検索用疑似単語の文字コードＳＣより小さく、最も文字コードＳＣに近い文字コードＶＣの固有疑似単語を有するレコード（以後、近似レコードと称す）を抽出する。

【００６５】そして、検索対象とする和文テーブルを１段下層のテーブル（例えば、和文最下層テーブルＪ２－１）とし、このテーブルの特定範囲に対して、上層のテーブルに対する場合と同様な検索処理が施される。ここで、特定範囲とは、上層のテーブル中の近似レコードに対応付けられた下層テーブル中のレコード、および、このレコードに後続する９９のレコードからなる。この検索処理が終了すると、処理はステップＳＢ５へ戻る。ステップＳＢ９は、ステップＳＢ５での判断が「Ｙｅｓ」となる場合の処理であり、検索された各レコードが有する各種アドレスを抽出する。

【００６６】検索されたレコードが和文テーブル（例えば、和文最上層テーブルＪ２－２）に存在する場合は、当該レコードに対応付けられた最下層の和文テーブル（例えば、和文最下層テーブルＪ２－１）中のレコードを抽出し、当該レコードに対応付けられる文字アドレスを抽出する。ここで抽出された最下層の和文テーブル中のレコードが有するサイズが複数である場合は、上記文字アドレスおよび後続する文字アドレス群から、順に、サイズの数だけ文字アドレスを抽出する。

【００６７】また、検索文字列が１文字である場合には、１文字テーブルＪ１中の抽出されたレコードに対応付けられる最下層の和文テーブル中のレコードを抽出する。この際、１文字テーブルＪ１中の検索されたレコードのサイズが複数であれば、当該レコードに対応付けられた和文最下層テーブルＪ２－１のレコードおよびこのレコードに後続するレコード群から、順に、サイズの数だけレコードを抽出する。こうして抽出された和文最下層テーブルＪ２－１中の各レコードに対応する文字アドレスを抽出し、昇順にソートする。

【００６８】次に、ステップＳＢ１０では、各検索用疑似単語に対応して抽出された文字アドレス群のうち、各検索用疑似単語間の距離に相当する差を有する文字アドレスの組を抽出し、抽出された組の先頭アドレスを抽出する。例えば、図１７に示すように、検索文字列が”経営危機”であれば、検索用疑似単語”経営”および”危機”間の距離は２である。

【００６９】したがって、検索用疑似単語”経営”に対応して抽出された文字アドレスと、検索用疑似単語”危機”に対応して抽出された文字アドレスとの差が２となる組を抽出する。この際、各検索用疑似単語に対応する文字アドレス群はソートされている為に、各々のアドレス群から小さい順に文字アドレスを抽出し、これらを比較することにより、差が２となる文字アドレスの組が容易に抽出される。そして、抽出された文字アドレスの組の先頭文字アドレス（”経営危機”の場合は’経’に対応する文字アドレス）が抽出される。

【００７０】また、検索文字列がアルファベットであり、例えば、欧文基本テーブルＥＢＴから１つあるいは複数のレコードＲ７が抽出された場合には、まず、当該レコードＲ７に対応付けられた仮想アドレステーブルＶＡＴ中のレコードＲ８を抽出する。そして、抽出されたレコードＲ８において、仮想アドレスのユニーク符号が同一のレコードＲ８について、ワード内アドレスの差が例えば、２となる仮想アドレスの組を抽出し、抽出された組の先頭文字アドレスを抽出する。

【００７１】ここで、例えば、検索用文字列が”ｗｏｒｌｄ”であれば、検索用疑似単語”ｗｏ”と”ｒｌ”との間隔は２、検索用疑似単語”ｒｌ”と”ｌｄ”との間隔は１である。したがって、検索用疑似単語”ｗｏ”に対応して抽出された仮想アドレス群と、検索用疑似単語”ｒｌ”に対応して抽出された仮想アドレス群とから、ユニーク符号が「Ａ」であり、かつ、ワード内アドレスの差が２となる仮想アドレスの組を抽出し、こうして抽出された仮想アドレスと、検索用疑似単語”ｌｄ”に対応して抽出され、ユニーク符号が「Ａ」である仮想アドレスとから、ワード内アドレスの差が１となる仮想アドレスの組を抽出する。

【００７２】そして、抽出された組の仮想アドレスのうち、先頭の仮想アドレス中のユニーク符号からワードリストＷＬ中のワードを特定する。特定されたワードには、ワードアドレスが対応付けられており、かつ、欧文範囲ＥＡの先頭ワードには、和文範囲ＪＡから連続する文字アドレス「３１７」も対応付けられているため、文書データ中における文字アドレスが得られる。もちろん、検索文字列が１文字である場合には、上述した連続性判断は行われない。

【００７３】次に、ステップＳＢ１１では、分脈意識検索か否かが判断される。この判断が「Ｙｅｓ」であればステップＳＢ１２へ、「Ｎｏ」であればステップＳＢ１３へ処理が進む。ステップＳＢ１２は、分脈意識検索で

ある場合の処理である。分脈意識検索であれば、検索文字列は複数（ここでは、説明を簡略化するために２つとする。以後、各検索文字列を第１の検索文字列、第２の検索文字列と称す）であり、ここでは、第１および第２の検索文字列に対する検索処理が終了したか否かが判断される。この判断が「Ｎｏ」であれば、ステップＳＢ３へ処理が戻り、未処理の検索文字列に対して上述した検索処理が施される。逆に、「Ｙｅｓ」であればステップＳＢ１３へ処理が進む。

【００７４】ステップＳＢ１３では、第１の検索文字列に対応して抽出される先頭文字アドレス群と第２の検索文字列に対応して抽出される先頭文字アドレス群とが比較され、両者の差が指定距離以下となる文字アドレスの組（以後、範囲内アドレス組と称す）を抽出する。この際、２つの文字アドレスで規定される文書データ中に、キャリッジリターン等の区切り記号が存在する場合には、両者の差が指定距離以下であっても、範囲内アドレス組から除外される。

【００７５】次に、ステップＳＢ１４では、範囲内アドレス組の数が１以上であるか否かが判断される。この判断が「Ｙｅｓ」であればステップＳＢ１５へ、「Ｎｏ」であればステップＳＢ１８へ処理が進む。ステップＳＢ１５では、順位指定があるか否かが判断される。この判断が「Ｙｅｓ」であればステップＳＢ１６へ、「Ｎｏ」であればステップＳＢ１７へ処理が進む。

【００７６】ステップＳＢ１６は、順位指定があった場合の処理であり、範囲内アドレス組内の文字アドレスの順序が、指定された順序と一致する組（以後、順序一致アドレス組と称す）を抽出する。ここで抽出される組が０でない場合には、処理はステップＳＢ１７へ進む。逆に、当該組が存在しない場合には処理はステップＳＢ１８へ進む。

【００７７】ステップＳＢ１７では、まず、各検索文字列に対応した先頭文字アドレス群に含まれる先頭文字アドレスの数に応じた表示データをディスプレイ９へ供給する。これにより、ディスプレイ９に表示されている検索メニューの出力フィールド１６に、抽出されたデータ数が表示される。これを視認したユーザーが、入力装置６を操作し、所定の指示データを検索処理装置７へ供給すると、当該装置７は、先頭文字アドレス群中の文字アドレスを有する文書データに応じた表示データをディスプレイ９へ供給する。

【００７８】こうして、図１６に示すように、検索結果がディスプレイ９上に表示される。この際、文書データ中の検索文字列に一致する文字列は、例えば、反転表示され、他の文字列と区別される。また、分脈意識モードであった場合には、指定範囲外あるいは順位が一致しなかった先頭文字アドレスの文字列に下線が付される。ここで、ユーザーは、入力装置６を操作し、他の検索結果等をディスプレイ９上に表示させる。

【００７９】また、ステップＳＢ１８は、ステップＳＢ６、ステップＳＢ１４、あるいはステップＳＢ１６において、検索対象文字列に一致する文字列を検索できなかったと判断された場合の処理であり、ディスプレイ９へ所定の表示データを供給し、「指定された条件の検索文字列は文書中に存在しませんでした」等のメッセージを表示させる。

【００８０】以上説明したように、本発明の一実施例によれば、文字列長が２の固有疑似単語を有する和文テーブルを階層的に構築する為に、和文テーブル自体のサイズを大きくすることなく、検索効率に優れた全文検索を行うことができる。また、１文字テーブルを設けた為に、文字列長が１の検索文字列に対する検索処理を迅速に行うことができる。さらに、ワードリストＷＬおよび欧文基本テーブルＥＢＴを作成した為に、各固有疑似単語に対応するサイズを適度な大きさとすることができ、検索効率を向上させることができる。

【００８１】また、欧文範囲ＥＡにおいて、ワード単位よりも小さな疑似単語単位での検索が可能になるために、語尾変化したワードを一度に検索することができる。例えば、入力装置６を介して”ｅｃｏｎ”という検索文字列を入力すると、”ｅｃｏｎｏｍｉｃ”，”ｅｃｏｎｏｍｙ”というワードを抽出することができる。

【００８２】なお、上述した一実施例においては、ＣＤ－ＲＯＭに全文データベースを記憶させる例を示したが、十分な記憶容量を有する記憶媒体であれば、ＣＤ－ＲＯＭでなくともよい。また、検索処理装置７はワークステーション等でも良く、パーソナルコンピュータである必要はない。さらに、１段下層のテーブルを分割する単位は９９である必要はなく、固有疑似単語の数に応じて設定することが望ましい。あるいは、検索作業をそのレコードで終了させるストップレコードを挿入するようにしてもよい。

【００８３】また、上述した一実施例においては、疑似単語の文字数を２文字として説明したが、２文字に限定されるものではなく、例えば、３文字、４文字、…というように複数文字であればよい。もちろん、疑似単語の文字数は、データベースの内容や検索処理の特徴等に応じて設定される。例えば、電子出版において、文書データが一般的な日本文である場合には、２文字程度に設定される。

【００８４】さらに、上述した一実施例では、電子出版に適用する例を示した為に、構築システム１が全文データベースの提供者に使用され、検索システム５が全文データベースのユーザーに使用されるように、それぞれ個別のシステムとして構成されるが、両者を一体のシステムとして構成し、電子出版以外の分野で用いられる一般的な全文データベースに対して適用可能であることは言うまでもない。

【００８５】

19

【発明の効果】以上説明したように、本発明によれば、アドレス付与手段が、文書データ中の各文字に連続するアドレスを付与し、アドレステーブル作成手段が、前記文書データ中の各文字と後続する文字とで構成される合計ｋ文字（ｋは２以上）の疑似単語を作成し、各疑似単語の先頭文字列に付与される各アドレスを、対応する疑似単語の文字コード順にアドレステーブルへ記憶する。そして、最下層構築手段が、固有の文字コードを有する疑似単語を固有疑似単語として前記インデックステーブル群中の最下層テーブルへ文字コード順に記憶するとともに、各固有疑似単語に前記アドレステーブル中の各アドレスを対応付ける。さらに、階層化手段が、前記インデックステーブル群中の最上層テーブルに記憶される固有疑似単語数が予め設定された数より大である場合、前記最上層テーブルを略均等に分割するように複数の固有疑似単語を抽出し、前記最上層テーブルの上層のテーブルへ前記複数の固有疑似単語を文字コード順に記憶する。インデックステーブル群は、このような階層構造を有するので、検索時において、十分な検索速度を得ることができるという効果がある。また、インデックステーブル群中の各テーブルに記憶される疑似単語は、固有の文字コードを有する固有疑似単語であるので、インデックステーブル群のサイズが小となるという効果を得ることができる。

【図面の簡単な説明】

【図１】本発明の一実施例による全文データベースシステムの概略構成を示すブロック図である。

【図２】同実施例による全文データベース構築処理の流れを示すフローチャートである。

【図３】同実施例で用いられる文書データの一例を示す図である。

【図４】マーク付加処理が行われた文書データの一例を示す図である。

20

＊【図５】各種アドレスが付与された文書データの一例を示す図である。

【図６】欧文対照テーブルＣＴの概略構成を示す図である。

【図７】ワードアドレステーブルＷＡＴおよびワードリストＷＬの概略構成を示す図である。

【図８】疑似単語テーブルＰＷＴの概略構成を示す図である。

【図９】仮想アドレステーブルＶＡＴおよび欧文基本テーブルＥＢＴの概略構成を示す図である。

【図１０】和文疑似単語テーブルＪＰＴの概略構成を示す図である。

【図１１】文字アドレステーブルＣＡＴおよび和文最下層テーブルＪ２－１の概略構成を示す図である。

【図１２】和文最上層テーブルＪ２－２および１文字テーブルＪ１等の概略構成を示す図である。

【図１３】本発明の一実施例による全文データベースシステムにおける検索処理の流れを示すフローチャートである。

【図１４】同システムにおける検索処理の流れを示すフローチャートである。

【図１５】検索メニューの一例を示す図である。

【図１６】検索結果の一例を示す図である。

【図１７】検索文字列の一例を示す図である。

【符号の説明】

３　　構築処理装置（アドレス付与手段、アドレステーブル作成手段、最下層構築手段、階層化手段）

６　　入力装置（入力手段）

７　　検索処理装置（分割手段、検索手段）

ＣＡＴ　　文字アドレステーブル（アドレステーブル）

Ｊ２－１　　和文最下層テーブル（最下層テーブル）

Ｊ２－２　　和文最上層テーブル（上層テーブル）

【図３】

多角経営で有名なＡ社（東京都）は、絶滅の危機にさらされている動物を保護する京都の団体に多額の寄付を行う一方、経常利益の１０％に相当する額の裏金を、Ｄ銀行経由でＥ国のペーパーカンパニーへ入金していた。１９９２年に、こうしたマネーロンダリングの事実が発覚すると、当時の全経営陣が失脚し、Ａ社は第２次オイルショック以来の経営危機に直面した。　JA

・・・

world wide economic help can help world economy ・・・　EA

【図６】

CT

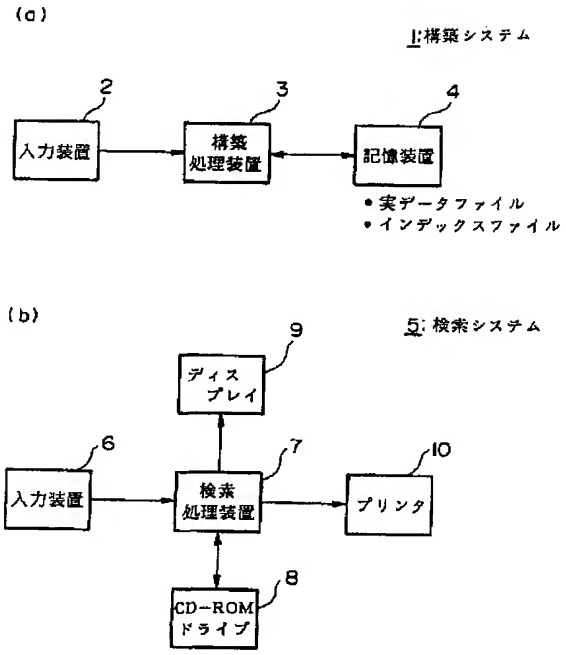
ワードアドレス	ワード
1	world
2	wide
3	economic
4	help
5	can
6	help
7	world
8	economy

RII　RII　RII

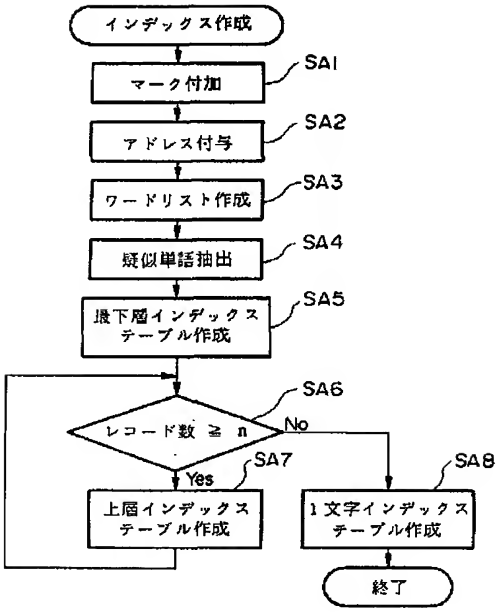
【図１７】

1 2 3 4
経営危機

【図１】



【図２】



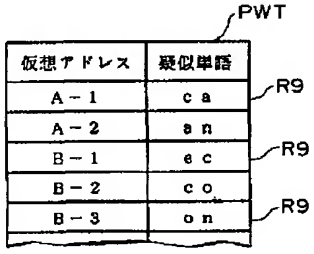
【図４】

多角経営で有名なＡ社（@東京都@）は、絶滅の危機にさらされている動物を保護する@京都@の団体に多額の寄付を行う一方、経常利益の１０％に相当する額の裏金を、Ｄ銀行経由でＥ国のペーパーカンパニーへ入金していた。１９９２年に、こうしたマネーロンダリングの事実が発覚すると、当時の全経営陣が失脚し、Ａ社は第２次オイルショック以来の経営危機に直面した。　JA

・・・

world wide economic help can help world economy ・・・　EA

【図８】



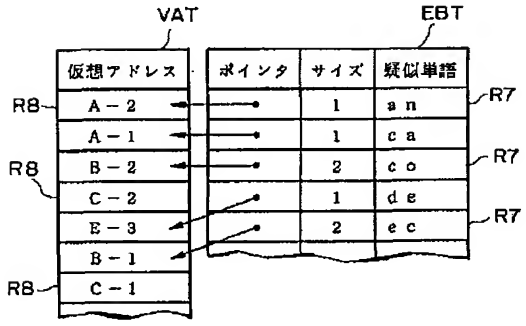
【図５】

1 2 3 4 5 6 7 8 9101112131415161718192021222324252627282930313 2
多角経営で有名なＡ社（@東京都@）は、絶滅の危機にさらされている
333435363738394041・・・
動物を保護する@京都@の団体に多額の寄付を行う一方、経常利益の１０％に相当する額の裏金を、Ｄ銀行経由でＥ国のペーパーカンパニーへ入金していた。１９９２年に、こうしたマネーロンダリングの事実が発覚すると、当時の全経営陣が失脚し、Ａ社は第２次オイルショック以来の経営危機に直面した。　JA

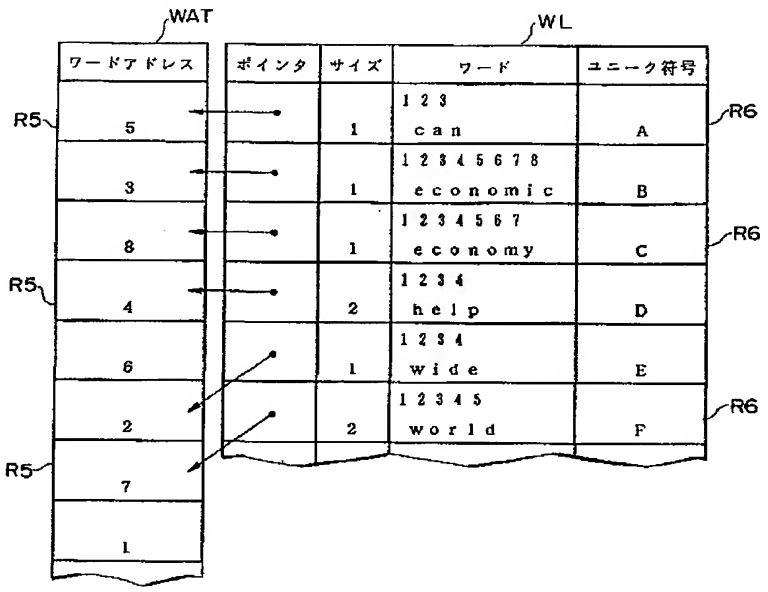
・・・

1(317)　2　3　4　5　6
world wide economic help can help　EA
7　8
world economy ・・・

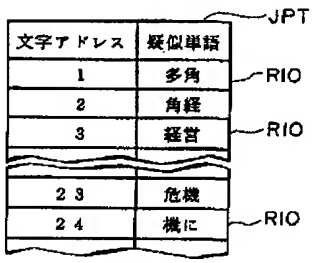
【図９】



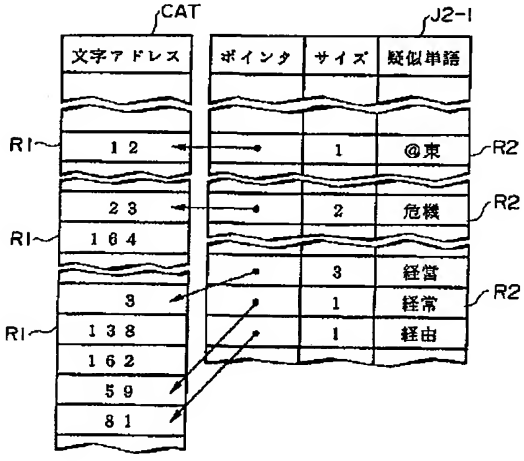
【図７】



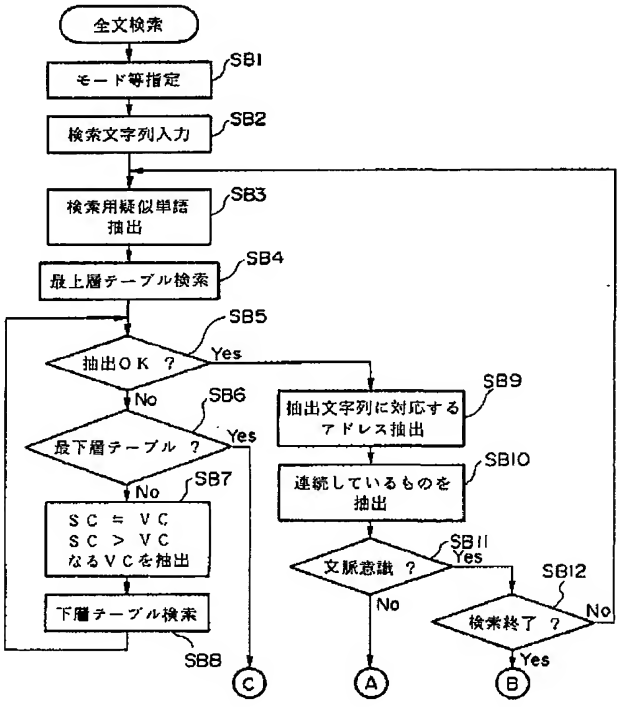
【図１０】



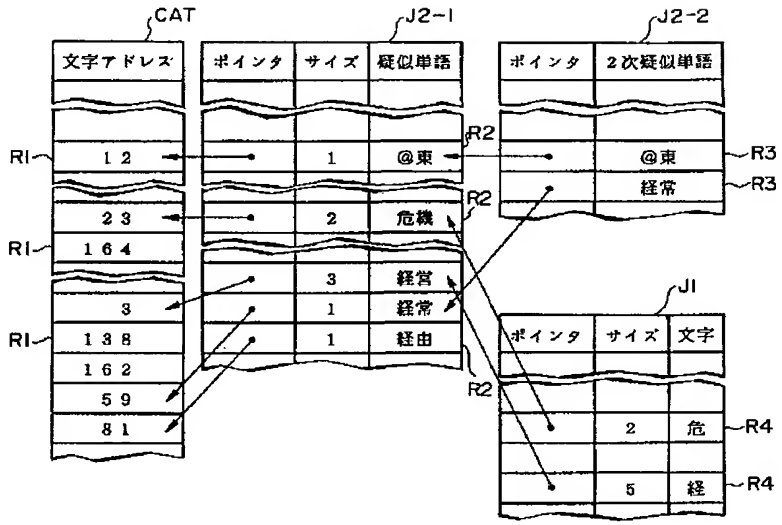
【図１１】



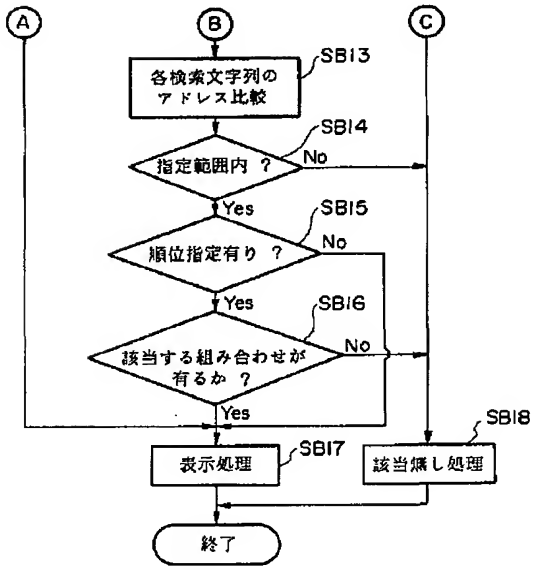
【図１３】



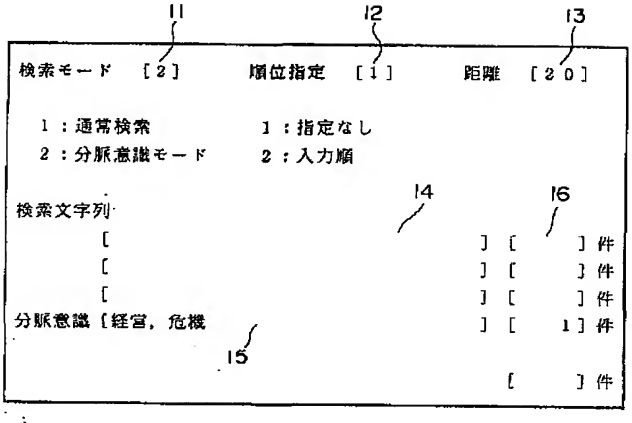
【図１２】



【図１４】



【図１５】



【図１６】

検　索　結　果　詳　細
多角経営で有名なＡ社（東京都）は、絶滅の危機にさらされている動物を保護する京都の団体に多額の寄付を行う一方、経常利益の１０％に相当する額の裏金を、Ｄ銀行経由でＥ国のペーパーカンパニーへ入金していた。１９９２年に、こうしたマネーロンダリングの事実が発覚すると、当時の全経営陣が失脚し、Ａ社は第２次オイルショック以来の[illegible]に直面した。・・・

フロントページの続き

(72)発明者　奈良　雅人
東京都台東区台東一丁目５番１号　凸版印刷株式会社内

(72)発明者　濱谷　群二
東京都台東区台東一丁目５番１号　凸版印刷株式会社内